**Problem Setting**

Reinforcement Learning with Rich Observations

- Rich contexts are generated from a small number of latent states;
- Agent only observes contexts rather than the latent states;
- Studied in e.g., [1, 2]

**Block Markov Decision Process**

- \( \mathcal{M}' := (\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathcal{r}, \mathcal{f}, H) \).
  - \( \mathcal{X} \) is a huge observation space.
  - \( \mathcal{S} \) is a small finite latent state space.
  - \( \mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{X}_s \), \( \mathcal{X}_s \cap \mathcal{X}_s' = \emptyset \).
  - \( \mathcal{X} \sim q(\cdot|s), f(x) = s, \text{ decoding function.} \)

**Numerical Test**

- A combination lock environment, hard for random exploration.
- \( H + 3 \)-dimensional observation = one-hot encoding of state + \( H \)-dimensional noise.

**Our Framework**

**Unsupervised Learning Oracle**

Given \( n \) samples \( \{x_i\}_{i=1}^n \) generated following \( \sum_{s \in \mathcal{S}} q(\cdot|s)\mu(s) \), with probability at least \( 1 - \delta \), \( \text{ULO} \) learns a function \( f: \mathcal{X} \rightarrow \mathcal{S} \) such that

\[
P_{x \sim \mu_\mathcal{X}}(f(x) = s) \geq 1 - g(n, \delta),
\]

where \( \lim_{n \to \infty} g(n, \delta) = 0 \).

**No-regret Tabular RL Algorithm**

For any MDP \( \mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{r}, H) \), \( \text{sf} \) runs for at most \( \text{poly}(|\mathcal{S}|, |\mathcal{A}|, H, 1/\varepsilon, \log(\delta^{-1})) \) episodes to learn an \( \varepsilon \)-optimal policy with probability at least \( 1 - \delta \). 

**Theoretical guarantee (informal version)**

**Theorem 1** Given an efficient \( \text{ULO} \) and a no-regret tabular RL algorithm, with at most \( \text{poly}(|\mathcal{S}|, |\mathcal{A}|, H, 1/\varepsilon, \log(\delta^{-1})) \) trajectories, we obtain an \( \varepsilon \)-optimal policy for the underlying BMDP with probability at least \( 1 - \delta \).

**Examples of ULO**

- Gaussian Mixture Models (GMM). In GMM, states lie in \( \mathbb{R}^d \) and observations are states plus some (truncated) zero-mean Gaussian noises.
- Bernoulli Mixture Models (BMM). In BMM, observations are points in \([0, 1]^d\). Every state \( s \) is frequency vector \( p^s \in [0, 1]^d \) such that \( q(x|s) = \prod_{i=1}^d (p_i^s)^{x_i} (1 - p_i^s)^{1-x_i} \).
- Subspace Clustering. In some cases, each state is a set of vectors and the corresponding observations are points lying in the subspace spanned by the state vectors.

Proper algorithms to serve as \( \text{ULO} \) can be found in literature.

**Conclusion**

We propose a provably efficient framework that turns an unsupervised learning algorithm and a no-regret tabular RL algorithm into an algorithm for RL problems with huge observation spaces.

**References**
